Predicting Bitcoin Price and Trend Forecasting Utilizing a Hybrid Approach of Financial and Textual Data

Final draft

Eleftherios Kanavakis
ekanavakis@ethz.ch
ETH Zurich
Switzerland

1 Introduction

This paper introduces a unique, integrative approach to Bitcoin (BTC) price and trend prediction by exploiting both financial and textual data. Bitcoin, as the leading cryptocurrency, plays a critical role in the financial market, making its price prediction an essential task. Traditional methods for predicting Bitcoin's price heavily rely on the analysis of historical financial data and trading indicators, such as moving averages. However, in an era characterized by information overflow and the widespread influence of social media on markets, it is reasonable to assume that public sentiment, as expressed through various social media platforms, has a significant impact on Bitcoin price fluctuations.

Consequently, this study proposes a hybrid prediction model that merges traditional financial indicators with a sentiment analysis of Bitcoin-related tweets. For the financial data, we leverage historical Bitcoin prices and trading indicators, which provides a quantitative perspective of market behavior. In contrast, for textual data, we utilize sentiment analysis on tweets related to Bitcoin to incorporate the qualitative aspect of public sentiment. This approach captures the essential factors contributing to the price changes in Bitcoin and improves the forecasting capability of the prediction model. By combining these two diverse but complementary sources of data, we aim to enhance the accuracy of BTC price and trend forecasting and to provide a more comprehensive understanding of the driving forces behind Bitcoin's market movements.

In our paper, we present insights from our feature diversification, model exploration efforts, and the impact of trading indicators. During a challenging bear market, we found that trading indicators significantly improved model performance, providing a remarkable 20% boost over the baseline and a notable 10% improvement over vanilla features for both trend and price prediction tasks. Furthermore, the addition of sentiment analysis from textual data (tweets) resulted in an additional 5% boost, aligning with related work in the field and emphasizing the synergistic potential of combining diverse data sources for cryptocurrency prediction.

In our model comparison, we found that XGBoost and MLP performed similarly in trend prediction, with LSTM demonstrating a slight advantage in handling complex temporal dependencies. For price prediction, both MLP and LSTM achieved lower RMSE values, suggesting their potential for continuous prediction tasks. However, all models encountered challenges due to the unpredictable nature of the test period, underscoring the importance of feature engineering and model selection in financial forecasting within dynamic markets.

In our study, we addressed a formidable challenge—a prolonged bear market period that presented heightened prediction difficulties compared to earlier studies in the literature. Our results highlighted a notable trend: the combination of trading indicators and the hybrid approach of incorporating tweets into our models proved instrumental. This strategy not only aligns with existing literature but also underscores the practical significance of integrating real-time sentiment from social media sources and valuable trading indicators. These combined efforts proved essential in enhancing prediction accuracy, particularly within the dynamic and ever-evolving landscape of financial markets.

2 Related Work

In the realm of financial prediction and sentiment analysis, a wealth of prior research has paved the way for our study. This section offers a concise overview of key literature and studies that have informed and inspired our approach. By examining the existing body of work, we gain valuable insights and context for our own research endeavors.

Ortu et al.[5] conducted a study combining financial and textual data to predict the trends of Bitcoin (BTC) and Ethereum (ETH) using data from 2017 to 2021. Their approach involved leveraging financial data similar to ours but gathering textual data from multiple sources, including Twitter, GitHub, and Reddit. Notably, their hybrid approach yielded a commendable F1 score of 0.83. A distinctive aspect of our study is its focus on a different time period marked by a bear market, offering a unique perspective on the challenges and opportunities associated with financial prediction during such conditions.

Tran et al.[7] conducted an extensive survey within the emerging field of predicting digital asset prices using Natural Language Processing (NLP). Their study provides a comprehensive overview of the evolving landscape of research and methodologies in this domain. By synthesizing insights from various sources, Tran et al. contribute to a deeper understanding of the challenges and advancements in predicting digital asset prices, providing valuable context for our own research in this dynamic field.

Sul et al.[6] employed a combination of tweets and financial data in their study to predict stock returns, achieving notable annual economic gains ranging from 11% to 15%. Their research highlights the potential synergy between social media sentiment expressed in tweets and financial data for generating substantial returns in the stock market. This achievement mirrors our own research's interest in leveraging social media sentiment, in our case, for predicting digital asset prices.

3 Dataset

The financial dataset that we are utilizing for this project originates from Bitfinex[2], one of the industry's leading cryptocurrency exchange platforms. The dataset covers a period from May 2018 to June 2023, equating to 1859 days or 44635 hourly candles. This time range is particularly interesting as it includes Bitcoin's recent bear market, the first to be significantly influenced by global macroeconomic factors since Bitcoin's inception, thus providing a rich context for our analysis. The dataset comprises the following key features:

- Open price: The opening price of Bitcoin at the start of each hourly candle.
- Close price: The closing price of Bitcoin at the end of each hourly candle.
- High: The maximum price Bitcoin reached within the duration of each hourly period.
- Low: The minimum price Bitcoin dipped to within each hourly interval.
- Date: The precise timestamp corresponding to each hourly candle.
- Volume BTC: The total volume of Bitcoin traded during each hourly period.
- Volume USD: The total volume of USD traded against the Bitcoin pair during each hourly period.

With the lowest recorded Bitcoin price at $3,215.2 and the highest soaring to $68,958, the dataset effectively illustrates the inherent volatility and unpredictability of the Bitcoin market. It is our belief that this comprehensive dataset, enhanced by the inclusion of a unique bear market scenario influenced by macroeconomic factors, can provide invaluable insights to develop a reliable model for Bitcoin price and trend prediction.

To complement our understanding, we've visually plotted1 the Bitcoin data over time, distinctly segregating the dataset into training and testing subsets, aiding in both our historical comprehension and future price and trend predictions.

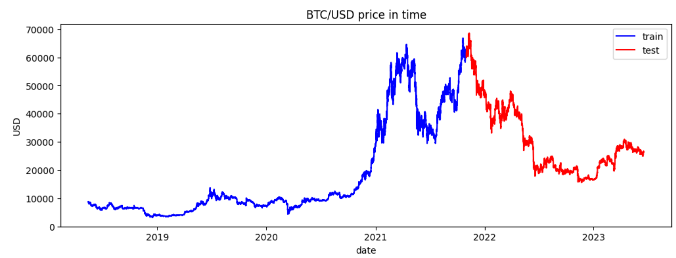

Figure 1. Bitcoin historical data in time

The textual data supplementing our financial dataset is derived from Twitter[1], an influential social media platform allowing users to post succinct messages or 'tweets'. Limited to 280 characters, these tweets encapsulate personal perspectives, news updates, opinions, and discussions on a broad spectrum of topics, including Bitcoin and other cryptocurrencies.

The Twitter dataset extends from February 5, 2021, to March 5, 2023. This duration enables us to glean a comprehensive overview of public sentiment towards Bitcoin over recent years. Encompassing a considerable count of 4,689,288 tweets, this dataset portrays a diverse range of opinions, dialogues, news, and sentiment concerning Bitcoin.

The data features fall into two broad categories:

User-Related Information: This includes aspects such as user_name, user_location, user_description, user_created, user_followers, user_friends, user_favourites, and user_verified, providing a broad context of the user profiles interacting with Bitcoin-related content.

Tweet-Related Information: This pertains to details like date, text, hashtags, source, and is_retweet, capturing the essence, timing, reach, and sentiment of the tweets.

Through the analysis of this rich dataset, we plan to assess the impact of public sentiment on Bitcoin's price and trend dynamics, supplementing the quantitative financial data with these qualitative insights.

4 Methodology

In this section, we adopt distinct modeling approaches for our two types of data: financial and textual. For the financial data, we utilize the historical price and volume trends of Bitcoin, while for the textual data, we perform sentiment analysis on relevant tweets. Finally, we merge these models to construct a comprehensive predictive model that capitalizes on both the quantitative insights from the financial data and the qualitative sentiment indicators from the Twitter data. This combined approach aims to offer a more holistic and accurate prediction of Bitcoin's price and market trends.

4.1 Financial Data Modelling

4.1.1 Tasks. In the context of the financial data, we will focus on two main tasks: Bitcoin (BTC) trend prediction and Bitcoin price prediction.

- Bitcoin Trend Prediction: This task is essentially a binary classification problem. Our goal here is to predict the close trend for timestep t+1, based on timestep t. In practical terms, if the closing price at timestep t is less than the closing price at timestep t+1, we consider it as an uptrend. Conversely, if the closing price at timestep t is greater than or equal to the closing price at timestep t+1, we interpret this as a downtrend.
- Bitcoin Price Prediction: The second task involves predicting the Bitcoin price, which is a regression problem. Here, the objective is to predict the closing price at timestep t+1, given the information available at timestep t.

To evaluate the performance of our models, we will utilize metrics inspired by related work in the field. For the task of trend prediction, we will apply the F1 score, a balanced measure of precision and recall. For price prediction, we will employ the Square Root of the Mean Squared Error (RMSE), a standard measure for regression tasks that assesses the average magnitude of prediction errors.

Both of these tasks contribute to our larger goal of developing a comprehensive predictive model for Bitcoin's price and market trends, taking into account not only these financial indicators but also the sentiment data extracted from Twitter.

4.1.2 Data Processing. In our data processing methodology and experiments, we employ three different strategies to build, compare, and improve our predictive models for Bitcoin price and trend.

Baseline: For our baseline, we utilize the "vanilla" or raw data as it is. This involves feeding the financial data, without any additional transformations or feature engineering, into our models to establish a benchmark for their performance.

Time-Specific Feature Addition: Building on our baseline, our second approach involves augmenting our vanilla data with time-specific features, such as the day of the week, month, and year. The rationale behind this strategy is grounded in the recurrent patterns often observed in financial markets, where certain behaviors tend to repeat on specific days, weeks, or months. Incorporating this temporal information could potentially enhance the learning capability of our models.

Combined Features: Our third and final strategy involves fusing the original features (technical indicators) with specific trading indicators. For the extraction and computation of these trading indicators, we leverage the Stockstats library. The technical indicators used include, but are not limited to, the following:

- Relative Strength Index (RSI): A momentum oscillator that measures the speed and change of price movements.
- Moving Average Convergence Divergence (MACD): A trend-following momentum indicator that shows the relationship between two moving averages of a security's price.
- Bollinger Bands: A volatility indicator that consists of a simple moving average (SMA) alongside upper and lower bands based on standard deviations.
- Average True Range (ATR): A technical analysis volatility indicator originally developed by J. Welles Wilder, Jr.
- Commodity Channel Index (CCI): A momentum-based oscillator used to help determine when an investment vehicle is reaching a condition of being overbought or oversold.

Through these three strategies, we aim to uncover the most effective approach for predicting Bitcoin's price and market trends, providing a comprehensive comparison and understanding of the different methods and their performances.

4.1.3 Modelling. In order to rigorously evaluate the effectiveness of our machine learning models for predicting trends in Bitcoin (BTC) prices, we first establish a baseline using a random classifier. This baseline provides a crucial point of reference, enabling us to assess how well our models distinguish significant trends from random noise, ultimately ensuring a comprehensive evaluation of our predictive capabilities.

To build the most robust predictive models, we'll be utilizing three varied architectures: XGBoost, Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM) networks. Before modeling, we ensure all input features are normalized to a similar scale, enhancing model training efficiency.

XGBoost, a gradient-boosting framework known for its effectiveness, will be applied as both a classifier and a regressor for trend and price prediction tasks respectively. Meanwhile, MLP, a type of artificial neural network, and LSTM, a form of recurrent neural network, will be utilized in both tasks. Configured as either classifiers or regressors as needed, they will work on predicting Bitcoin's market trend and price. These models collectively provide a comprehensive framework for tackling our prediction tasks.

4.2 Sentiment Analysis

4.2.1 Data Processing. For the textual data sourced from Twitter, we apply specific filtering criteria to streamline the dataset and focus on the most relevant information for our task. Our initial step involves retaining only those tweets that contain Bitcoin-related hashtags such as 'btc', 'bitcoin', or 'Bitcoin'. This not only enables us to hone in on the most pertinent discussions and sentiments about Bitcoin but also helps in managing the vast volume of data, given our processing resources.

In addition to hashtag filtering, we implement two additional preprocessing steps to streamline our dataset and focus on the most relevant information for our analysis. Firstly, we filter users with more than 100 followers and who have a description to ensure the legitimacy of the accounts and prioritize those with more comprehensive profiles, reducing the potential for noise in our dataset. Secondly, we exclude tweets that are retweets, effectively eliminating redundant information and ensuring that our analysis is based on unique content. These preprocessing steps collectively enhance the quality and relevance of the data for our analysis.

Additionally, we restrict our dataset to include only tweets that originate from mobile devices. This decision is motivated by two key reasons. Firstly, mobile-originated tweets are often less prone to being generated by automated bots, thereby reducing the noise in our data. Secondly, this filter, much like the hashtag criteria, contributes to our data pruning efforts. Through these measures, we ensure our textual data is both manageable and highly relevant to Bitcoin trend and price prediction.

Having pared down our initial dataset to a manageable volume of relevant tweets, we now turn our attention to the text processing phase. This stage is essential in preparing our data for the subsequent modeling. Here we refine the content of the tweets, a process integral to the efficacy of our sentiment analysis model.

Initially, we discarded the hashtag symbol (#) to focus solely on the textual substance of the tags. Following that, we employed regular expressions to remove hyperlinks from the tweets, as these do not contribute to the overall sentiment conveyed by the text. Finally, we extricated all mentions from the tweets, which are typically usernames prefixed with '@'. This allows our model to focus on the key text content without being swayed by user-specific data. These meticulous cleaning steps resulted in a curated dataset of 1,140,256 tweets, primed for effective sentiment analysis.

4.2.2 Modelling. For sentiment analysis, we utilize FinBERT[3], a specialized NLP model, plays a pivotal role in our analysis of financial sentiment within Bitcoin-related tweets. It evolves from BERT through intensive training on a rich financial corpus, notably the 'Financial PhraseBank' dataset by Malo et al. (2014). Powered by the transformer architecture, it excels in capturing the intricate linguistic nuances of finance. In our technical journey through cryptocurrency discourse, FinBERT[3] serves as our analytical anchor, deciphering sentiment patterns with precision.

In our process, we feed the tweet text data into this fine-tuned model, which subsequently outputs a sentiment score and a label for each tweet. Potential sentiment labels are Negative,Positive and Neutral. This method provides us with a potent mechanism for gleaning sentiment insights from a substantial corpus of tweet data, thereby enriching our understanding of public sentiment around Bitcoin during the study period.

4.3 Combination of methods

To seamlessly integrate the sentiment analysis results with our forecasting models, we developed a comprehensive process for harmonizing these two distinct data streams. Our strategy revolves around computing the hourly average sentiment score for each Bitcoin-related tweet, thereby yielding a consolidated sentiment score for each hour within our study period. In addition to this average sentiment score, we have augmented our final dataset with three additional features: the count of negative, positive, and neutral sentiments per hour. This holistic fusion of financial and sentiment data empowers our forecasting models to harness the collective wisdom of both quantitative and qualitative insights, potentially leading to more informed and accurate predictions.

At this point, it is important to mention that, in scenarios where sentiment scores are absent or where a robust prediction cannot be made, we rely on the neutral label assigned by FinBERT[3]. This strategic decision ensures continuity in our dataset, preventing interruptions caused by missing or unreliable sentiment data. With FinBERT's inherent capability to assign a neutral sentiment label, our models seamlessly process and analyze the entirety of our dataset, even in instances where sentiment analysis proves challenging.

5 Results

5.1 Financial Data

In this section, we present the results obtained from our analysis of financial data, on the tasks of trend prediction and price forecasting for Bitcoin. These results are critical as they shed light on the effectiveness of both our data processing and chosen models in predicting Bitcoin's future behavior, based on its historical financial data.

Building on these observations from the results table (Table 2), it becomes evident that our comprehensive efforts in diversifying feature sets and exploring various models have yielded valuable insights. Firstly, we found that time-specific features did not significantly enhance our models' predictive performances. This outcome is particularly notable given the challenging test period marked by a pronounced bear market. However, it's worth highlighting that trading indicators, on the other hand, made a substantial impact. They improved our models' performance by a remarkable 20% compared to the baseline and by 10% compared to vanilla features, for both trend and price prediction tasks.

When it comes to model comparison, XGBoost and MLP presented comparable performances for the trend prediction,

Data	Model	Trend(F1)	Price(RMSE)
Vanilla	Baseline	0.49	
Vanilla	XGBoost	0.49	821.51
Vanilla	MLP	0.49	238.98
Vanilla	LSTM	0.54	255.75
Vanilla + Time	XGBoost	0.49	828.77
Vanilla + Time	MLP	0.49	238.98
Vanilla + Time	LSTM	0.54	255.75
Vanilla + Trading	XGBoost	0.54	609.36
Vanilla + Trading	MLP	0.54	238.49
Vanilla + Trading	LSTM	0.61	489.99

Table 1. Results for trend prediction and price prediction tasks using different data experiments across various models.

whereas LSTM slightly pulled ahead, indicating its superior capability in managing complex temporal dependencies. In the price prediction task, MLP and LSTM managed to achieve lower RMSE values, suggesting their potential edge in handling continuous prediction tasks. However, all models were equally challenged by the unpredictable nature of the test period, a testament to the fundamental difficulties involved in financial forecasting.

To complement our understanding, we've visually plotted2 the Bitcoin data over time, and the results from our forecasting model.

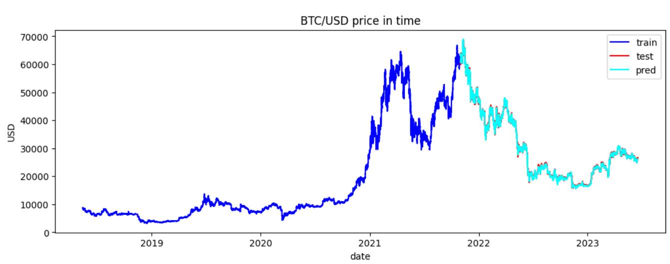

Figure 2. Bitcoin price forecasting results

5.2 Financial and Textual Data

In this section, we present the results of our hybrid approach, which combines sentiment analysis on tweets with historical price data for Bitcoin price prediction. It is important to note that the dataset utilized in this study exhibits a slight variation from the initial description. Specifically, the tweet data we collected spans from 2021 to 2023, prompting us to prune the Bitcoin value dataset to ensure alignment with our tweet data timeframe. However, it is worth highlighting that, while a test set does indeed exist, it covers almost the entire period of the pruned dataset. Consequently, we refrain from using this test set for evaluation purposes, as it lacks the necessary separation in time to effectively assess the model's predictive performance. Instead, we opted for a 5-fold cross-validation approach to test the hypothesis of whether incorporating textual data could enhance the predictive capabilities of our method. In addition to the aforementioned considerations, it's essential to note that, for the same practical reasons, we did not include Long Short-Term Memory (LSTM) models in our experiments. These decisions were made in light of the project's constrained scope, which limited our resources and time allocation for processing an extended period of textual data. Despite these practical considerations and necessary shortcuts, our primary aim was to rigorously evaluate the impact of textual data on our model's predictive accuracy, a question we address comprehensively in the ensuing sections.

Data	Model	Trend(F1)	Price(RMSE)
Vanilla	Baseline	0.49	
Vanilla	XGBoost	0.51	1365
Vanilla	MLP	0.51	2078
Vanilla + Text	XGBoost	0.53	1365
Vanilla + Text	MLP	0.52	2000

Table 2. Results for trend prediction and price prediction tasks using different data experiments across various models.

The presentation of our results reveals a notable pattern: the inclusion of textual data consistently enhances the performance of our model, yielding improvements of up to 5% in both trend and price prediction tasks. This consistent performance boost, validated through our rigorous 5-fold cross-validation approach, not only underscores the efficacy of our hybrid methodology but also aligns with prior literature and our initial assumptions. It reaffirms the growing body of evidence in the field, highlighting the valuable insights that can be extracted from the amalgamation of sentiment analysis on social media data and historical price information when predicting Bitcoin's behavior in the market. This alignment with existing research further strengthens the credibility of our findings and reinforces the potential of leveraging textual data for improved cryptocurrency price prediction.

6 Conclusion

In conclusion, while our study sheds light on predicting financial trends and prices during challenging bear markets, it's important to note some limitations. We didn't fine-tune FinBERT for our specific tweet dataset, a potential avenue for performance improvement, as discussed here[4]. Additionally, incorporating Bitcoin halving as an exogenous feature could enhance predictions, given its impact on prices. Exploring time series transformers is another promising direction for improving our models. Finally, in future endeavors, it is crucial to expand our text data processing to encompass the entire Bitcoin dataset and predict sentiment for each hourly candle. Additionally, diversifying our textual data sources can enhance dataset richness. As the cryptocurrency landscape evolves, these steps hold promise for more accurate price predictions, signifying the importance of further research in this area. In conclusion, this study paves the way for future advancements in leveraging textual data for cryptocurrency forecasting.

References

[1] [n.d.]. Bitcoin Tweets. https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweet

[2] [n.d.]. BTC/USD Historical Data. https://www.cryptodatadownload.com/data/bitfinex/

[3] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv:1908.10063 [cs.CL]

[4] Jingwei Ni, Zhijing Jin, Qian Wang, Mrinmaya Sachan, and Markus Leippold. 2023. When Does Aggregating Multiple Skills with Multi-Task Learning Work? A Case Study in Financial NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 7465–7488. https://doi.org/10.18653/v1/2023.acl-long.412

[5] Marco Ortu, Nicola Uras, Claudio Conversano, Silvia Bartolucci, and Giuseppe Destefanis. 2022. On technical trading and social media indicators for cryptocurrency price classification through deep learning. *Expert Systems with Applications* 198 (2022), 116804. https://doi.org/10.1016/j.eswa.2022.116804

[6] Hong Sul, Alan Dennis, and Lingyao Yuan. 2016. Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns: Trading on Twitter. *Decision Sciences* 48 (06 2016). https://doi.org/10.1111/deci.12229

[7] Trang Tran. 2022. Predicting Digital Asset Prices using Natural Language Processing: a survey. arXiv:2212.00726 [cs.CY]